

Searching for a particle of unknown mass and lifetime in the presence of an unknown non-monotonic background

Mike Williams

Massachusetts Institute of Technology, Cambridge, MA, United States

ABSTRACT: Many extensions to the Standard Model of particle physics hypothesize the existence of new low-mass particles. Typically there are few theoretical constraints on the mass or lifetime of such particles. This requires the experimentalist to perform a search in which both the mass and lifetime of the particle are unknown. Such searches for low-mass particles are complicated by the possible presence of resonances and other non-monotonic backgrounds. This paper presents a simple and fast approach to assigning significance and setting limits in such searches.

Contents

1. Introduction	1
2. Search Strategy	3
2.1 Mass	3
2.2 Lifetime	6
3. p-Values	7
4. Limits	9
5. Extensions	11
6. Summary & Discussion	12
A. Profile Likelihood	17
B. Resonances	19
C. Two-Sample Non-Parametric Goodness-of-Fit Tests	23
D. Lifetime Test Results	24
E. Fast Algorithm	26

1. Introduction

Many extensions to the Standard Model (SM) of particle physics hypothesize the existence of new low-mass particles. Typically there are few theoretical constraints on the mass or lifetime of such particles. Therefore, the experimentalist is required to perform a search in which both the mass and lifetime of the particle are unknown. Such searches for low-mass particles are complicated by the possible presence of resonances and other non-monotonic backgrounds whose probability density functions (PDFs) are not well known or constrained.

The null results from numerous searches for new particles that decay into leptons has shifted the focus of many theorists towards *leptophobic* bosons. For example, Tulin [1] proposes a force that couples weakly to quarks at the QCD scale. Experimentally, this requires a search for a MeV or GeV mass boson that decays to $\pi^+\pi^-\pi^0$. If such a boson decays promptly¹, then searching for

¹Throughout this paper I refer to decays as either: (prompt) the separation of the decay point from the production point is too small to be resolved by the detector and (displaced) this separation is resolvable.

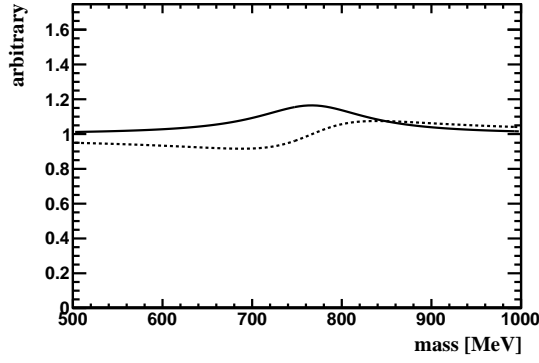


Figure 1. Simulated PDF including interference between $B \rightarrow K^* \mu \mu$ and $B \rightarrow K^* \rho(\mu \mu)$ for two arbitrary phase differences between the two amplitudes (shown as solid and dashed lines).

it requires dealing with an irreducible resonance background. It is preferable to perform this search with the data blinded; however, the reaction under study is governed by non-perturbative QCD and so the only *a priori* assumption about the SM PDF that can be made is: any resonance that couples to the same final state as the new boson may contribute to the SM background. The number of known resonances is large and new resonances are still being discovered, even those that only contain light quarks [2]. Furthermore, parametrization of resonances is reaction-dependent and interference effects are often significant. Therefore, the number of nuisance parameters in the background PDF is very large and there will likely always be some component to the PDF that is simply not modeled accurately to better than a few percent. Even in dedicated amplitude analyses performed without blinding the data, it is often difficult to evaluate the systematic uncertainties associated with the fit PDF due to the large number of nuisance parameters. It is also difficult to obtain a fit model which describes the data at the percent level for all masses.

Searches for new bosons decaying to leptons are not immune to the problem of resonances. Consider, *e.g.*, the rare decay $B \rightarrow K^* \mu \mu$ which is an excellent laboratory to search for a new low-mass boson that couples to mass [3]. This decay can have a contribution from $B \rightarrow K^* \rho(\mu \mu)$. Using the published LHCb $B \rightarrow K^* \mu \mu$ results [4] and the ratio of branching fractions $\mathcal{B}(B \rightarrow K^* \rho) \mathcal{B}(\rho \rightarrow \mu \mu) / \mathcal{B}(B \rightarrow K^* \mu \mu)$ from the PDG [5] one obtains a predicted yield of $B \rightarrow K^* \rho(\mu \mu)$ that is less than one event; however, this calculation ignores interference. Figure 1 shows a toy prediction obtained using the measured branching fractions $\mathcal{B}(B \rightarrow K^* \mu \mu)$, $\mathcal{B}(B \rightarrow K^* \rho)$ and $\mathcal{B}(\rho \rightarrow \mu \mu)$, and the assumption that the full amplitudes for $B \rightarrow K^* \mu \mu$ and $B \rightarrow K^* \rho(\mu \mu)$ interfere². Naively using the ratio of branching fractions suggests $B \rightarrow K^* \rho(\mu \mu)$ is negligible, but the interference cross term could be large enough to generate a “local” 5% effect. In principle the ρ contribution could be parametrized (with some uncertainty); however, the branching fractions of many resonance decays to $\mu^+ \mu^-$ and of $B \rightarrow K^*$ resonance have yet to be measured and so such contributions would be unconstrained. Since the effect generated can also depend on interference, a full model allowing all resonances to interfere (which requires introducing unknown relative phases for each resonance) must be constructed.

²I ignore muon spin states here which produces an overestimate of the size of this interference effect, but for illustrative purposes this model is sufficient.

For the case of displaced decays, the largest background contributions are often due to some mis-reconstructed candidates. These include material interactions, photon conversions, *etc.* Given that the searches considered in this paper must scan a large mass and lifetime range, predicting (blindly) the full list of displaced backgrounds and parametrizing them will often not be possible. Furthermore, resonances may contribute to the displaced search via decays of beauty or charm particles. *E.g.*, in the aforementioned $B \rightarrow K^* \mu \mu$ search, it is possible to obtain the K^* and $\mu^+ \mu^-$ from two different B decays at LHCb. This could result in a candidate that contains what appears to be a long-lived charmonium resonance. It is desirable to approach such a search by making as few assumptions as possible about the backgrounds.

This paper presents a simple approach for performing a search for a new low-mass particle of unknown mass and lifetime. The method is described in Sec. 2. Obtaining global p -values, including the so-called trials factor or look elsewhere effect, is discussed in Sec. 3, while setting upper limits and the coverage properties of such limits is discussed in Sec. 4. Discussion and summarizing are provided in Sec. 5.

2. Search Strategy

The strategy that will be employed is to perform a scan in mass since the new particle mass is not known. The step size of the scan will be $\sigma(m)/2$, where $\sigma(m)$ is the detector mass resolution. For each test mass value a search will be conducted for evidence of a particle with no constraints on its lifetime. For illustrative purposes, the following toy model is used:

- the searched region in mass is 1 GeV wide³;
- the mass resolution is taken to be $\sigma(m) = 2$ MeV, giving a step size of 1 MeV;
- the expected number of prompt events (in the absence of signal) is 1000;
- the expected number of displaced events (in the absence of signal) is 100;
- both prompt and displaced events are generated uniformly in mass.

This toy data set is sufficient for illustrating how the proposed method works. Some variations are considered later in the text.

2.1 Mass

As an alternative to fitting the mass distribution in data with a PDF whose accuracy is *a priori* unknown and difficult to validate, I propose the following simple approach: For each test mass value in the scan, $m(\text{test})$, use the mass sidebands to estimate the expected background⁴ yield. The signal region (where signal events would be observed if a new particle of mass close to $m(\text{test})$)

³I just take the mass region searched to be 0 to 1000 MeV since the absolute mass values do not matter; *i.e.*, shifting the mass window to a range that starts at some allowed kinematic limit has no affect on applying this method. For this toy study I also do not specify what the decay products of the particle being searched for are since these also do not matter apart from identifying which resonances may contribute to the data.

⁴Any candidate that does not come from the decay of a new particle is considered as a background in the search.

exists) is defined as $|m(\text{test}) - m| < 2\sigma(m)$, while the background region is the sidebands defined as $3\sigma(m) < |m(\text{test}) - m| < (2x + 3)\sigma(m)$, where x is the ratio of the size of the signal to sideband regions (these regions do not need to be the same size). The factors that contribute to choosing x are discussed at the end of this subsection.

If the background PDF is approximately linear in the region $|m(\text{test}) - m| < (2x + 3)\sigma(m)$, then the observed yield in the background region provides an estimate of the expected yield in the signal region. The presence of resonances and some types of background will violate this so-called local-linear approximation; this is dealt with below. Under the local-linear assumption, the likelihood is given by

$$L(n_s, n_b | s, b) = \mathcal{P}(n_s, s + b) \cdot \mathcal{P}(n_b, xb), \quad (2.1)$$

where \mathcal{P} denotes the Poisson PDF, $n_{(s,b)}$ denote the yields in the (signal, background) regions and s and b are the signal and background rates in the signal region. It is straightforward to account for uncertainty in the relationship between the sideband and search window background yields, or equivalently in deviations from local linearity in the background PDF, by augmenting the likelihood as follows:

$$L(n_s, n_b, x | s, b, y) = \mathcal{P}(n_s, s + b) \cdot \mathcal{P}(n_b, yb) \cdot \mathcal{G}(y, x, \sigma_y), \quad (2.2)$$

where \mathcal{G} denotes the Gaussian PDF and σ_y is the uncertainty on the scaling factor between the background yield in the signal and background regions⁵. If one can estimate the size of possible deviations from local linearity, *e.g.*, due to resonance contributions, then these can be incorporated into the likelihood via σ_y . The profile likelihood can then be used to obtain the significance of any excess of events (see, *e.g.*, Ref. [6]) and/or to set limits [7]. See Appendix A for a more detailed discussion on the likelihood.

In the previous paragraph the nominal background PDF is taken to be linear but in principle any background PDF can be used. For a low-statistics search the linear approximation will be sufficient; however, for cases where the sample sizes are large, then the uncertainty due to deviations from local-linearity may be large compared to the statistical uncertainties. All that is required to use this method is that based on the observed yields in the sidebands, an estimate for the expected yield in the signal region can be obtained. If a non-linear background PDF is chosen as nominal, then x represents the scaling factor between the expected background yields in the sidebands and signal region, which may no longer be simply the ratio of the size of the regions. The σ_y term is again the uncertainty in the scale factor (independent of how x is defined). See Sec. 5 for discussion on non-linear background PDFs. For the remainder of the method overview, I will assume a nominal background PDF of local linear.

Appendix B provides a study of the deviation from local linearity due to resonance contributions. These deviations are a function of the width of the resonance, Γ , the fraction of the total PDF near $m(\text{test})$ due to the resonance, and of $\sigma(m)$ and x . The conclusions are as follows:

- for $\Gamma/\sigma(m) > 20$, even if the resonance makes up close to 100% of the PDF the choice $x = 1$ is still local-linear to about 10% (wide resonances are *safe*);

⁵This likelihood can be used for any value of n_b ; however, Appendix A shows that for the case where n_b is large the background region Poisson term can be replaced with a Gaussian one which results in a much faster algorithm.

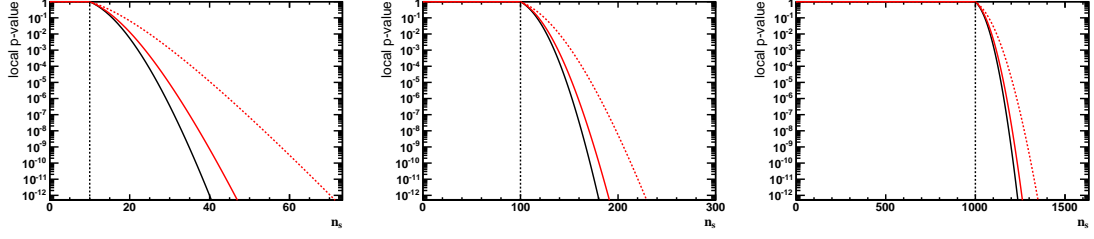


Figure 2. The p -value vs n_s for (left) $b = 10$, (middle) $b = 100$, (right) $b = 1000$ for (solid black) b known with no uncertainty, (dotted red) $x = 1$ and (solid red) $x = 5$.

- for $\Gamma/\sigma(m) < 5$, the deviation from local-linear is large unless the resonance contribution is small (narrow resonances must be vetoed);
- for $5 < \Gamma/\sigma(m) < 20$, the local-linear approximation is valid at the 10% level for moderate resonance contributions, but not valid if the resonance is dominant.

Assuming a $\mathcal{O}(10\%)$ value is chosen for σ_y , then wide resonances can effectively be ignored (including those that have yet to be discovered). If nothing is known about the possible size of a contribution from a narrow resonance, then the region near the resonance mass should be vetoed. If some limits can be placed on the size of a resonance contribution, then this veto may not be required. Such limits must be determined in each analysis independently.

The key point is that narrow resonances are typically well known (including their branching fractions to many final states), whereas wide resonances are not well known. The sideband approach allows the analyst to ignore wide resonances by accounting for their non-linear effects via the σ_y term in the likelihood. Narrow resonances likely must be vetoed, but there are few of these and their properties are well measured. Intermediate-width resonances are also accounted for automatically by σ_y provided they do not dominate the local PDF. Such cases should be rare and can be studied using alternative decays of the resonance or via the data directly using bins $\mathcal{O}(10\sigma(m))$ wide.

Other categories of background that have a broad peaking structure are also handled naturally in this approach. The study in Appendix B can be applied to non-resonant backgrounds with the same conclusions drawn: only a background whose peak is narrow relative to $\sigma(m)$ must be vetoed; all other backgrounds are accounted for via σ_y (the analyst does not need to know what these are or account for them individually). *E.g.*, in the $B \rightarrow K^* \mu \mu$ search, partially reconstructed charm particle decays can be ignored, while any $J/\psi \rightarrow \mu^+ \mu^-$ contribution must be vetoed.

The parameter x should be optimized for each analysis. The larger x is chosen to be, the less statistical uncertainty there is on the background rate; however, this also increases the size of the region in which local-linearity is assumed. Figure 2 shows a comparison of using sidebands to estimate b to when b is known (which I assume here is not possible; this is added for illustrative purposes to show the best possible performance which could be obtained if b could be known). Setting $\sigma_y/x = 0.1$ typically loses very little power (at most 10-20% sensitivity to the signal rate relative to the ideal situation of a known background PDF), so a good rule of thumb would be to choose x as large as possible such that $\sigma_y/x = 0.1$ is still valid.

2.2 Lifetime

I now proceed to considering lifetime, τ , information and assume that the background can be factorized into two components: (1) a prompt background where the signal candidate's children all originate from the same spatial point as the rest of the final state and (2) a displaced background where they do not. The lifetime PDF for type (1) is given by the detector resolution. The lifetime PDF for type (2) is assumed to be unknown. In principle it may be obtainable from the mass sidebands; however, looking at these sidebands is forbidden in a blind analysis when the mass is unknown. Furthermore, in many cases the displaced background data will be sparse making obtaining a reliable estimate of its PDF (which likely depends on m) impossible even if the data is not blinded.

If the lifetime distribution of both prompt and displaced backgrounds is the same in the signal and background regions, then a non-parametric two-sample goodness-of-fit test can be used to test the null hypothesis at each $m(\text{test})$. Appendix C discusses several such distribution-free tests, while Appendix D shows the results of applying these to the toy data set described above. The conclusion of these studies is that such tests are far from optimal and require introducing the assumption that the displaced background τ PDF does not vary with m (which may not be true and would be difficult to validate).

An alternative (and simple) approach is to define two τ regions at each $m(\text{test})$: (1) a prompt region (*e.g.*, $\tau < 3\sigma(\tau)$) and (2) a displaced region ($\tau > 3\sigma(\tau)$). The mass sidebands in each region can be used to estimate the expected background rate. The joint likelihood is defined as

$$L(n_s^{\text{prompt}}, n_b^{\text{prompt}}, n_s^{\text{displ}}, n_b^{\text{displ}}, x | s^{\text{prompt}}, b^{\text{prompt}}, y^{\text{prompt}}, s^{\text{displ}}, b^{\text{displ}}, y^{\text{displ}}) = \quad (2.3)$$

$$L(n_s^{\text{prompt}}, n_b^{\text{prompt}}, x | s^{\text{prompt}}, b^{\text{prompt}}, y^{\text{prompt}}) \cdot L(n_s^{\text{displ}}, n_b^{\text{displ}}, x | s^{\text{displ}}, b^{\text{displ}}, y^{\text{displ}}),$$

i.e., the likelihood is the product of the likelihoods from each region individually (each obtained in the same manner as discussed in the previous section). The profile likelihood (Λ), which is the ratio of the maximum likelihood with the signal rate fixed to the maximum likelihood with the signal rate free to vary (see Appendix A), for this two-region test is the product of the profile likelihoods from each region. This is true because no assumption is made about the lifetime of the new particle and so there is no assumed relationship about the signal yields in each region. The asymptotic distribution of $-2\log\Lambda$ is a χ^2 with two DOFs. Appendix D shows that this simple approach, which only uses the lifetime information to determine which region each candidate falls in, is nearly optimal except when $\tau \sim \sigma(\tau)$.

The approach presented in this paper only assumes that variation in the lifetime distribution *vs* mass is slow enough that the linear approximation holds in both the prompt and displaced regions. The possible deviations from linearity are accounted for by σ_y . In this paper I use a constant σ_y in the notation but σ_y is allowed to depend on m and different values of σ_y can be used in the prompt and displaced regions. This is also true of x : it can be chosen to be different values for different test masses and in the prompt and displaced regions. Using variable σ_y and x does not introduce any additional complexity.

Finally, I note that one can run both the two-region profile likelihood test and a two-sample test. Appendix D shows that this approach provides a small increase in performance in the region near $\tau \sim \sigma(\tau)$; however, there is an important assumption required to use the two-sample test. This

approach requires that all lifetime PDFs are the same in the signal and background regions, *i.e.*, that locally the τ PDF is independent of m . This may not be true for displaced backgrounds and would be difficult to validate unless the number of displaced-background candidates is large. Given that the gain in performance is small and the additional complexity introduced into the analysis is non-negligible, I conclude that unless one has a reason to expect $\tau \sim \sigma(\tau)$ and a method for validating the τ PDF m dependence, that the two-region profile likelihood test is the best option.

While it may be surprising to the reader that such a simple approach performs so well, it is often the case that analysis-specific information provides great power. In this case, factorizing the background into two categories using the known detector resolution is key to this search. Note that this procedure allows the analyst to completely ignore the m and τ distributions for any background that does not form a narrow peak relative to $\sigma(m)$. One instead relies on the fact that such backgrounds will populate the signal and sideband regions and that locally their PDFs are approximately linear. The possible deviation from linearity is incorporated into the likelihood via σ_y . Narrow peaking backgrounds, *e.g.*, $J/\psi \rightarrow \mu\mu$ in the $B \rightarrow K^* \mu\mu$ search, must be vetoed in each lifetime region unless it can be shown (or known) that they can only contribute to one region. Such backgrounds should be few and easy to identify.

3. p -Values

The full procedure involves first determining the local p -values at each $m(\text{test})$, then obtaining the global p -value of the most significant excess observed in the full mass range. An outline of the procedure is as follows:

- The full mass range is scanned in steps of $\sigma(m)/2$.
- An independent test is performed for each $m(\text{test})$ where the signal and background regions are defined as $|m - m(\text{test})| < 2\sigma(m)$ and $3\sigma(m) < |m(\text{test}) - m| < (2x + 3)\sigma(m)$, respectively (x should be optimized for each analysis).
- The signal and background regions are divided into prompt and displaced sub-regions. The quantities $n_s^{\text{prompt}}, n_b^{\text{prompt}}, n_s^{\text{displ}}, n_b^{\text{displ}}, x$ are the inputs to the local two-region likelihood.
- The profile likelihood provides the local test statistic (see Appendix A). Since the lifetime of the new particle is unknown, there is no assumed relationship between the signal rate in the prompt and displaced regions.
- For each $m(\text{test})$ an approximate local p -value is obtained using the asymptotic formula for the profile likelihood test statistic. As discussed below, the accuracy of the asymptotic formula for this test only needs to be good enough to properly select the most significant local excess.
- The minimum p -value from the full mass scan is selected as the test statistic to which a significance is to be assigned. This approach is motivated by the assumption that there is at most one new particle contributing to the data sample.

The global p -value is not the minimum local p -value as this would ignore the so-called trials factor, or look elsewhere effect, induced by the fact that a large number of tests have been performed.

Appendix E discusses the fact that this test has been designed such that it can be run on 10M data sets for $\mathcal{O}(1000)$ $m(\text{test})$ values in about 2 hours on a single CPU core. This permits determining the global p -value without the need for using asymptotic formulae. The only reliance on asymptotic formulae is in selecting the minimum local p -value; therefore, the accuracy of the asymptotic formula only needs to be sufficient to properly select the most significant local excess. There is no need to interpret the local p -values as probabilities under the null hypothesis.

Figure 3 shows the local p -values obtained from a single simulated toy data set. Since the step size in $m(\text{test})$ is smaller than the signal and sideband regions, neighboring tests are correlated. This produces a jagged-looking distribution. The test mass near 520 has the minimum p -value so would be selected to be assigned a significance in this data set. To obtain the significance the procedure is as follows:

- Obtain the minimum local p -value from the data as described above.
- Get an approximate null PDF from the data by ignoring the region near the most significant excess and obtaining a smooth PDF from the remaining data (interpolating into the removed region). Below I show that the details of how this is done are not important.
- Generate an ensemble of simulated data sets from the PDF from the previous bullet point. The global p -value is the fraction of simulated data sets that have a minimum local p -value less than that observed in the data.
- The number of generated data sets determines the statistical uncertainty on the global p -value. The most likely outcome, no evidence for a new particle, will require only $\mathcal{O}(100)$ data sets. To confirm $> 3\sigma$ requires $\mathcal{O}(1000)$ while $> 5\sigma$ requires $\mathcal{O}(10^7)$.
- Confirming $> 5\sigma$ can be done on a single CPU core on a laptop. In the unlikely (and exciting) event that not a single simulated data set in 10^8 has a minimum p -value less than that of the data, the asymptotic formula from Ref. [8] can be used to obtain an approximate significance (if one is desired).

Figure 4 shows the cumulative distribution of minimum local p -values obtained from 10M toy data sets. One can see that to obtain a global 3σ in this example requires a local p -value of about e^{-13} . The trials factor then is $\mathcal{O}(400)$ which is roughly the width of the full mass range divided by $\sigma(m)$ (that is 500 in this example). The asymptotic distribution provides an underestimate of the significance in this example for small p -values due to the small sample sizes used in each local test. Note that the true PDF used in the toy model is local-linear and so setting $\sigma_y/y = 0.1$ is an overestimate of the local non-linearity. In this example, such an overestimate produces only a minor shift (towards lower significance; it produces a conservative estimate).

Figure 5 shows that the cumulative p -value distribution obtained using an alternate (highly non-monotonic) PDF but the same sample size and detector resolution (and, thus, test mass values). There is very little dependence on the data PDF. This means that the exact PDF used to generate the pseudo data sets is not important. For example, one could simply bin the data in a histogram with

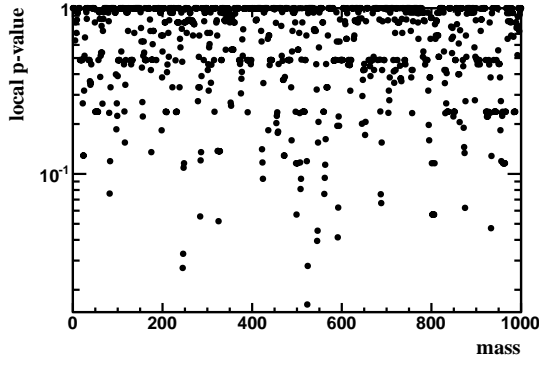


Figure 3. Local p -values obtained from a single simulated data set ($x = 1$).

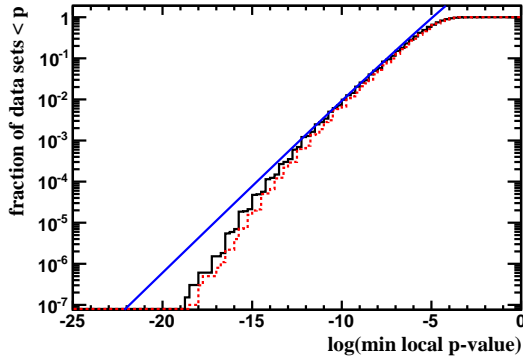


Figure 4. Cumulative distribution of minimum local p -values obtained using simulated (black) toy-model events compared to the (blue) asymptotic expectation[8]. Two variations of the test are shown (both use $x = 1$): (solid) no uncertainty in the relationship between the signal and background regions and (dashed red) $\sigma_y/y = 0.1$. The discrepancy at very small p -values between the asymptotic and solid distributions is due to low statistics in each local test region. The data sets were generated using a PDF that is local linear so it is expected that using $\sigma_y/y = 0.1$ is an overestimate of the scaling uncertainty which results in an underestimate of the significance.

wide bins (relative to $\sigma(m)$), remove the most-significant excess region, and use spline interpolation to obtain a background PDF (which interpolates into the removed most-significant region). This will be accurate enough to produce a reliable cumulative p -value distribution.

To summarize: One can confirm up to $> 5\sigma$ without using asymptotic formulae in about an hour on a laptop. Assignment of a significance beyond this level can be done using the asymptotic formula as an estimate (if this is desired). Figure 5 demonstrates that even an oscillatory PDF is handled naturally (I did not input any knowledge of this PDF to the method except that $\sigma_y/y = 0.1$) provided the features of the PDF are wide relative to $\sigma(m)$.

4. Limits

Upper limits are to be set as a function of m and τ for all m . In the event that a globally significant excess is observed, the analyst can additionally perform PDF-based fits to determine estimators for

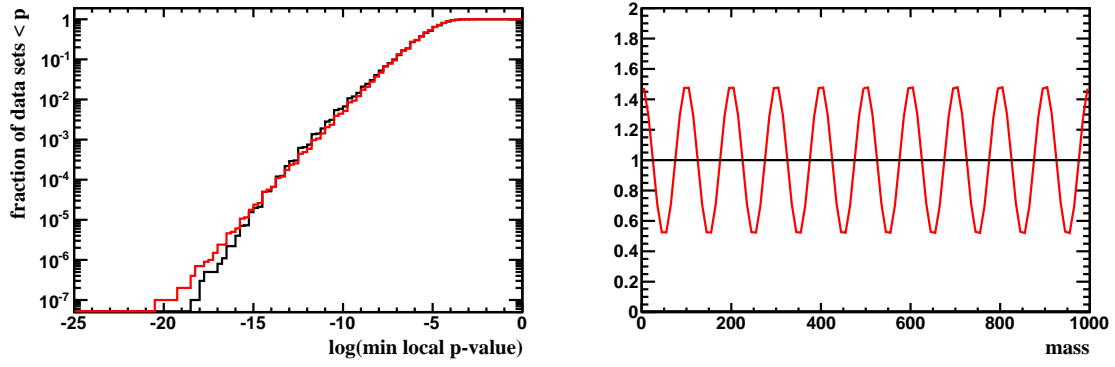


Figure 5. Cumulative distribution of minimum local p -values obtained using simulated toy-model events distributed according to the PDFs shown at right (black is the nominal linear model, while red is a non-monotonic alternative). The distribution of local p -values has little dependence on how the data is distributed as expected. This distribution is driven by the size of the mass range being searched and $\sigma(m)$. Both results here are shown for $\sigma_y/y = 0.1$ and $x = 1$.

the new particle mass and lifetime. The two-region profile likelihood can be used to set the limits after making the following modifications to the likelihood function:

- for each value of τ , there is a relationship between the number of signal events expected in the prompt and displaced regions;
- a Gaussian term is added to the likelihood to account for uncertainty (due to detector efficiency) in the fraction of signal expected in the prompt and displaced regions;
- another Gaussian term is added to the likelihood to account for uncertainty in the absolute detector efficiency scale of the signal (most likely relative to a normalization decay mode).

The likelihood for each $(m(\text{test}), \tau)$ is then given by

$$\begin{aligned}
 L(n_s^{\text{prompt}}, n_b^{\text{prompt}}, n_s^{\text{displ}}, n_b^{\text{displ}}, x, \tau | \dots) &= L(n_s^{\text{prompt}}, n_b^{\text{prompt}}, x | \varepsilon \cdot s \cdot f, b^{\text{prompt}}, y^{\text{prompt}}) \\
 &\times L(n_s^{\text{displ}}, n_b^{\text{displ}}, x | \varepsilon \cdot s \cdot (1 - f), b^{\text{displ}}, y^{\text{displ}}) \\
 &\times \mathcal{G}(f, f_{\text{MC}}(\tau), \sigma(f)) \times \mathcal{G}(\varepsilon, \varepsilon_{\text{MC}}(\tau), \sigma(\varepsilon)), \quad (4.1)
 \end{aligned}$$

where f is the fraction of signal events in the prompt region with expected value from simulation $f_{\text{MC}}(\tau)$ and uncertainty $\sigma(f)$, and ε is the efficiency (typically relative to some normalization reaction) with expected value from simulation $\varepsilon_{\text{MC}}(\tau)$ and uncertainty $\sigma(\varepsilon)$. The limits are then set by scanning the profile likelihood using the same method, including the handling of special circumstances, discussed in detail in Ref. [7] but using the likelihood given in Eq. 4.1⁶. For setting limits there is no need to generate 10M toy data sets; thus, I do not provide analytic solutions and, instead, use MINUIT to numerically scan the profile likelihood.

In the toy analysis I choose to normalize the new particle yield to the observed prompt yield in the full mass region. This emulates the situation where the prompt background is dominantly

⁶This method actually returns a confidence interval whose lower limit may be > 0 . Since the significance is discussed above, here I only study upper limits but the method will produce a lower limit as well.

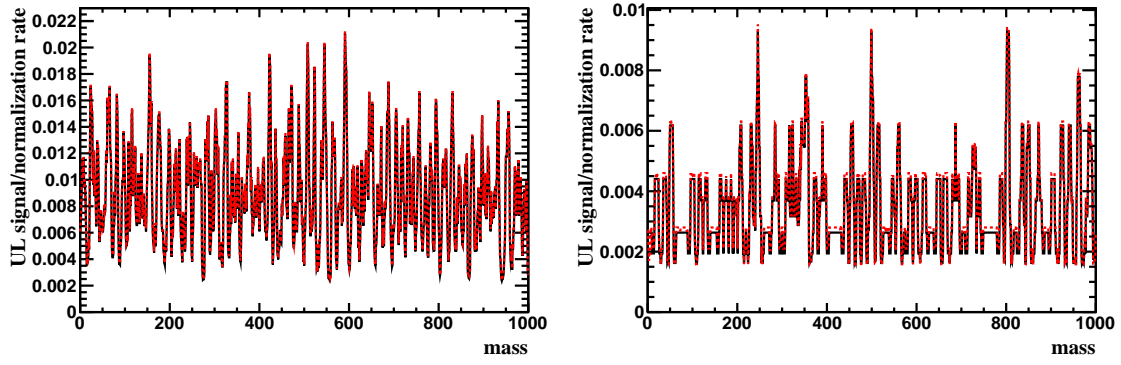


Figure 6. Comparison between limits obtained for a single simulated data set from (black) the TRolke class (only Poisson uncertainties) and (dashed red) the method discussed in this work (with $\sigma_y/y = 0.1, x = 1$) for (left) test $\tau = 0$ and (right) $\tau = 1000\sigma(\tau)$. In the left plot only the prompt region is used with TRolke, while in the right plot only the displaced region is used (TRolke only handles one region). For very small and large τ , using only one region is an excellent approximation to the full method presented here; thus, the agreement between TRolke and this method for these τ values is expected.

a well-known SM process. I take the ratio of the efficiency for detecting the new particle to the normalization process to be one. In reality there will be some dependence of ε on m and τ and there will be some candidates in the prompt region that do not come from the normalization mode, but these just scale the limits (and contribute to $\sigma(\varepsilon)$) so I will not discuss them here.

For the cases where the test τ is $\ll \sigma(\tau)$ or $\gg \sigma(\tau)$, to a good approximation only the prompt or displaced region matters. Therefore, in such cases the one-region results obtained using the TRolke class in ROOT [9] should be close to those produced here provided σ_y/y is small. Figure 6 shows that this is the case. With $\sigma_y/y = 0.1$, the limits returned by this method are slightly larger than TRolke (which here takes the uncertainty on b to be purely Poisson). If $\sigma_y = 0$ then for small or large τ the limits produced by this method are the same as TRolke.

Figure 7 shows how the limits depend on τ for a single toy-model data set. Since in this example $n_b^{\text{prompt}} > n_b^{\text{displ}}$, the limits decrease with increasing τ . Figure 8 shows the dependence of the limits on τ expected (obtained from the 10M toy data sets generated). For the toy-model data, whose PDF is uniform in m , the expected limits have no m dependence. In general this will not be the case.

Figure 9 shows the coverage obtained for various configurations of this method and various background rates expected in the prompt and displaced regions. The coverage properties are good except for at small s . At small s the method over covers but this is expected and unavoidable. Otherwise, the method tends to over cover by only a few percent. Note that the small over coverage shown in Fig. 9 is due to the very low statistics of the samples studied. The possible observations are discrete, and with such low statistics it is not possible to obtain perfect coverage.

5. Extensions

The method is not restricted for use where the local-linear approximation is applicable. For example, in high-statistics searches, one could consider using multiple sideband regions for each m (test)

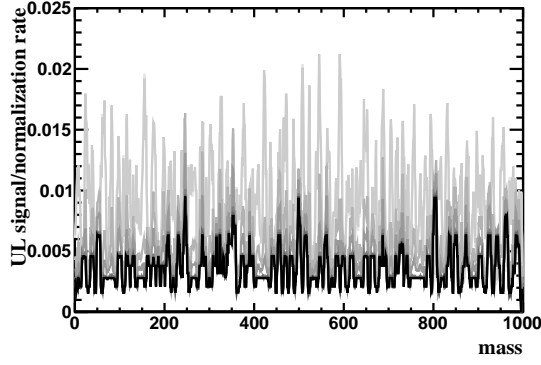


Figure 7. Upper limits from a single data set on the ratio of new particle production to the SM reaction for τ values (lighter to darker) $\tau = (0, 1, 5, 10, 50, 100, 500, 1000)\sigma(\tau)$.

and using spline interpolation (of whatever order is sufficiently constrained) instead of the local-linear approximation. Another approach could be to unblind some small fraction of the data and obtain estimates for the background PDF there. However the background estimates are obtained for each $m(\text{test})$, they will have some uncertainty σ_y and so this method can still be applied using Eq. A.10 in Appendix A. There is no restriction to any particular local background shape.

6. Summary & Discussion

This paper presents a simple likelihood-based approach for searching for a particle of unknown mass and lifetime in the presence of unknown non-monotonic backgrounds. Instead of exhaustively fitting the data with background PDFs containing hundreds of nuisance parameters, I propose to use the local-linear (or alternative local PDF) assumption and simple sideband counting. Deviations from the nominal local PDF are parametrized via a single parameter σ_y in the likelihood. If σ_y is overestimated, then only minor underestimation of the p -values is found. This allows the analyst to concentrate on a small number of possible narrow peaking backgrounds and effectively ignore all other contributions.

The lifetime information is used in a simple two-region approach which is nearly optimal except when $\tau \sim \sigma(\tau)$. This permits the avoidance of attempting to parametrize displaced background PDFs from a few sparse observed data. This method is fast enough to verify significance $> 5\sigma$ in an hour on a laptop; thus, reliance on asymptotic formulae is not required. Furthermore, when setting limits only the integrated detector efficiency in each lifetime region is required to be determined for each mass. A detailed determination of the uncertainty on the detector efficiency vs lifetime is not required since the τ information is only used to classify candidates as prompt or displaced. For limit setting, uncertainties on the integrated detector efficiency and on the fraction of events that fall in each region are included in the likelihood and the coverage is shown to be good.

Finally, Fig. 10 shows the low-recoil (high $M(\mu\mu)$) data observed by LHCb in the decay $B \rightarrow K\mu\mu$ [14]. There is a clear and sizable contribution from the $\psi(4160)$ resonance. The size of this contribution was unexpected. What would have happened if a blind analysis of this data had been performed to search for a new prompt particle that decays to $\mu\mu$? If a fit using a monotonic

background PDF and no $\psi(4160)$ term had been used⁷, then I estimate that the local p -value near 4200 would have given a local significance of $4 - 5\sigma$. Using the method presented in this paper (in only the prompt region), I estimate that the p -value is about 0.3(0.4) for $\sigma_y = 0(0.1)$ for $x = 1$. This is due to the fact that $\Gamma(\psi(4160)) \sim 10\sigma(m)$ and so locally the resonant peak only deviates from local-linearity by about 15%. No false claim of a discovery would have been made. Furthermore, given the large number of nuisance parameters and sample size, a fit-based approach would not provide much (possibly no) additional sensitivity to new particles (no matter how much effort was put into developing the fit model).

Acknowledgments

I thank Tim Gershon for useful comments that helped improve this paper. This work was supported by US NSF grant PHY-1306550.

⁷This is unlikely to have happened since prior to observing this data one would still have expected some charmonium contributions. The more likely scenario would be that a fit would have been performed that contained every ψ state with masses and widths free to vary within their nominal values; various other resonant shape parameters free to vary; and the relative phase of each amplitude free. This would then have required a serious systematic study to determine the p -values. Here I am simply using this as an example of how an unexpectedly large wide resonance contribution is handled naturally in the method presented in this paper.

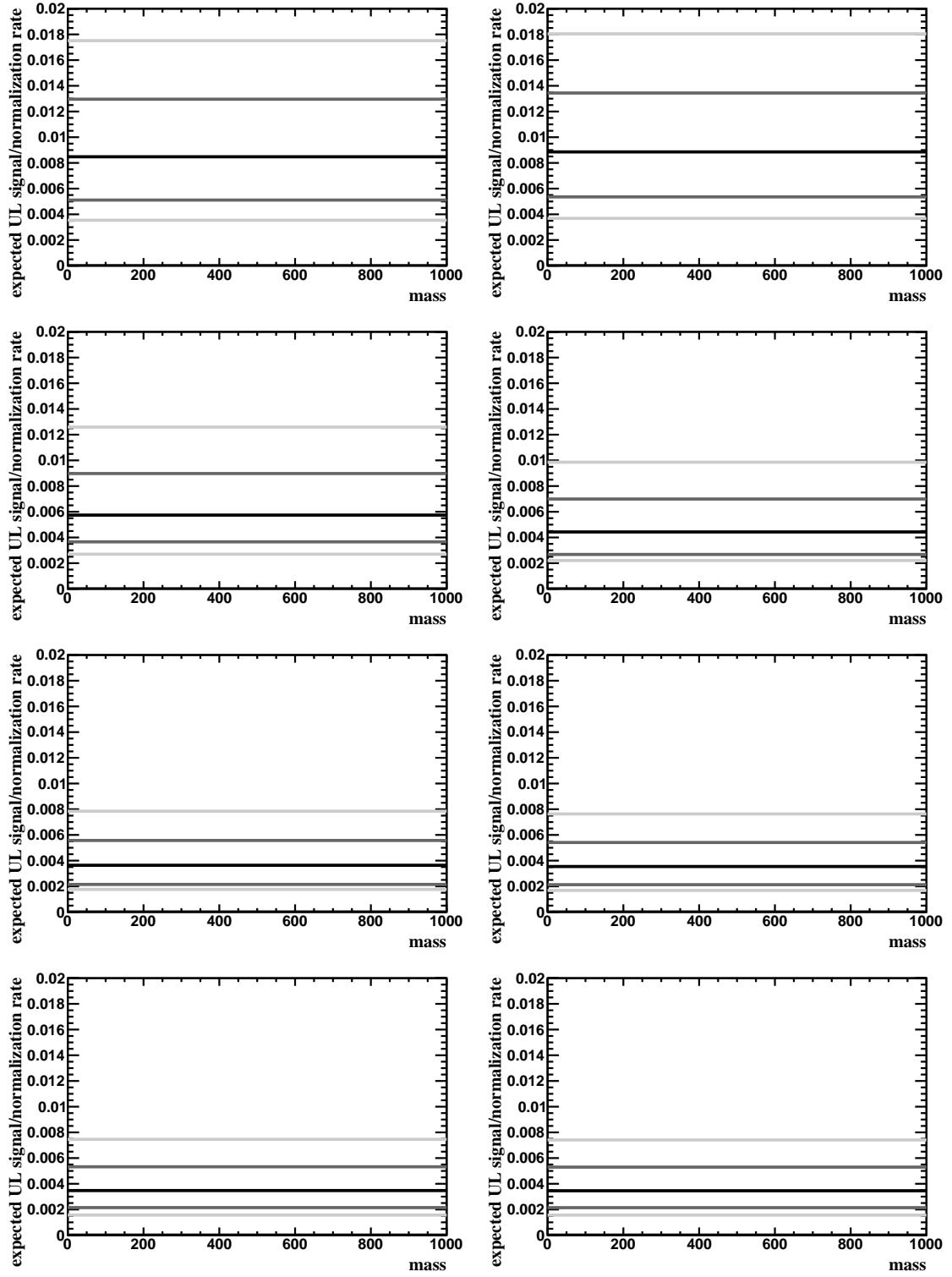


Figure 8. Expected upper limits on the signal rate relative to the normalization mode *vs* m for simulated toy-model data ($\sigma_y/y = 0.1, x = 1$). The lines are (black) mean, (dark gray) enclose the 1σ and (light gray) 2σ intervals. From left to right, top to bottom $\tau = (0, 1, 5, 10, 50, 100, 500, 1000)\sigma(\tau)$.

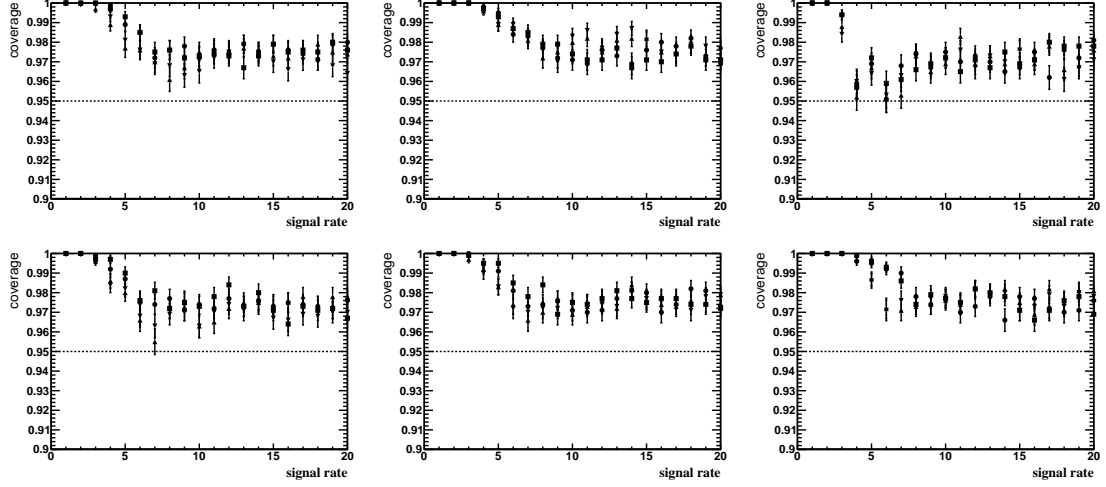


Figure 9. Coverage obtained using (circles) $x = 1, \sigma_y = 0$, (squares) $x = 1, \sigma_y = 0.1$, (up triangles) $x = 10, \sigma_y = 0$, (down triangles) $x = 10, \sigma_y = 0.1$ for (top) $b^{\text{prompt}} = 10, b^{\text{displ}} = 0.1$ and (bottom) $b^{\text{prompt}} = 10, b^{\text{displ}} = 10$ and (left) $\tau = 0$, (middle) $\tau = \sigma(\tau)$ and (right) $\tau = 10\sigma(\tau)$. The dashed line shows the desired 95% coverage. This method typically over covers by a few percent for the low-statistics cases shown here.

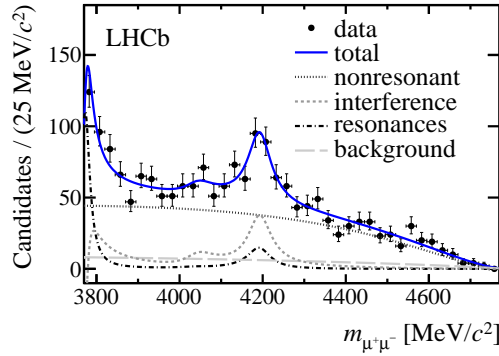


Figure 10. LHCb $B \rightarrow K\mu\mu$ data observed at low recoil (image taken from Ref. [14]). The peaking structure is due to the $\psi(4160)$.

References

- [1] S. Tulin, *New weakly coupled forces hidden in low-energy QCD*, Phys.Rev. **D89**, 114008 (2014).
- [2] COMPASS Collaboration, *Observation of a new narrow axial-vector meson $a_1(1420)$* , [arxiv:1501.05732].
- [3] M. Freytsis, Z. Ligeti and J. Thaler, *Constraining the axion portal with $B \rightarrow K \ell^+ \ell^-$* , Phys.Rev. **D81**, 034001 (2010).
- [4] LHCb Collaboration, JHEP **1308**, 131 (2013). [arXiv:1304.6325]
- [5] J. Beringer et al. (Particle Data Group), Phys.Rev. **D86**, 010001 (2012).
- [6] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur.Phys.J. **C71**, 1554 (2011).
- [7] W.A. Rolke, A.M. Lopez and J. Conrad, *Limits and confidence intervals in the presence of nuisance parameters*, Nucl.Instrum.Meth. **A551**, 493-503 (2005).
- [8] E. Gross and O. Vitells, *Trial factors or the look elsewhere effect in high energy physics*, Eur.Phys.J. **C70**, 525 (2010) [arXiv:1005.1891].
- [9] The ROOT data analysis framework: `root.cern.ch`
- [10] A. Kolmogorov, G.Ist.Ital.Attuari **4**, 83-91(1933); N. Smirnov, Annals of Mathematical Statistics **19**, 279-281 (1948).
- [11] H. Cramer, *On the composition of elementary errors*, Scandinavian Actuarial Journal (1928); R.E. von Mises, *Statistik und Wahrheit*, Julius Springer (1928).
- [12] T.W. Anderson and D.A. Darling, Ann.Math.Statist. **23**, 193-212 (1952); J.Am.Stat.Assoc. **49**, 268 (1954).
- [13] R.A. Fisher, *Statistical methods for research workers*, Oliver and Boyd Ltd., London (1925).
- [14] LHCb Collaboration, Phys.Rev.Lett. **111**, 112003 (2013).

A. Profile Likelihood

The Poisson likelihood for observing n_s events in the signal region and n_b events in the background region is

$$L(n_s, n_b | s, b) = \mathcal{P}(n_s, s + b) \cdot \mathcal{P}(n_b, xb), \quad (\text{A.1})$$

where the Poisson PDF is defined as

$$\mathcal{P}(n, \lambda) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad (\text{A.2})$$

and s is the signal rate, b is the background rate and x is the ratio of the size of the signal and background regions. The profile likelihood is then defined as

$$\Lambda(s | n_s, n_b) = \frac{L(s, \hat{b}(s) | n_s, n_b)}{L(\hat{s}, \hat{b} | n_s, n_b)}, \quad (\text{A.3})$$

where $\hat{b}(s)$ denotes the value of b that maximizes the likelihood for fixed s , and \hat{s}, \hat{b} maximize L in general. These parameters can be obtained analytically⁸ and are $(\hat{s}, \hat{b}) = (n_s - n_b/x, n_b/x)$ and $\hat{b}(0) = (n_s + n_b)/(1 + x)$. In general, the estimator for b for any value of s is

$$\hat{b}(s) = \left(n_s + n_b - (1 + x)s + \sqrt{(n_s + n_b - (1 + x)s)^2 + 4(1 + x)sn_b} \right) / 2(1 + x). \quad (\text{A.4})$$

Asymptotically, $-2 \log \Lambda$ behaves as a χ^2 with 1 DOF and so an approximate p -value can be obtained from Λ . *N.b.*, one may worry about the possibility that when $s < 0$ $\hat{b}(s)$ could become imaginary. See discussion at the end of this appendix on how the $s < 0$ case is dealt with.

It is straightforward to account for uncertainty in the relationship between the sideband and search window yields by augmenting the likelihood as follows:

$$L(n_s, n_b, x | s, b, y) = \mathcal{P}(n_s, s + b) \cdot \mathcal{P}(n_b, yb) \cdot \mathcal{G}(y, x, \sigma_y), \quad (\text{A.5})$$

where the Gaussian PDF is defined as

$$\mathcal{G}(z, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\mu)^2/2\sigma^2}. \quad (\text{A.6})$$

Now y is the scale factor that relates the yields in the sideband and search regions whose PDF is taken to be a Gaussian with mean x and uncertainty of σ_y . Following the same approach to maximizing L produces an algebraically intractable set of three equations. Making the approximation that $\sigma_y/y < 1/\sqrt{n_b}$, then to leading order in σ_y/y

$$\hat{s} \approx n_s - n_b/x \quad (\text{A.7})$$

$$\hat{y} \approx x + \sigma_y^2 (n_b/x - \hat{b}(\hat{s}, \sigma_y = 0)) \quad (\text{A.8})$$

$$\hat{b} \approx \hat{b}(\hat{s}, \sigma_y = 0, x \rightarrow \hat{y}), \quad (\text{A.9})$$

where $\hat{b}(\hat{s}, \sigma_y = 0)$ is the result given in Eq. A.4 and $\hat{b}(\hat{s}, \sigma_y = 0, x \rightarrow \hat{y})$ uses the same equation but replaces x with \hat{y} everywhere. Figure 11 shows that this approximation is accurate out to p -values of about $\mathcal{O}(10^{-12})$ when the relative uncertainty on the scale factor is smaller than the relative

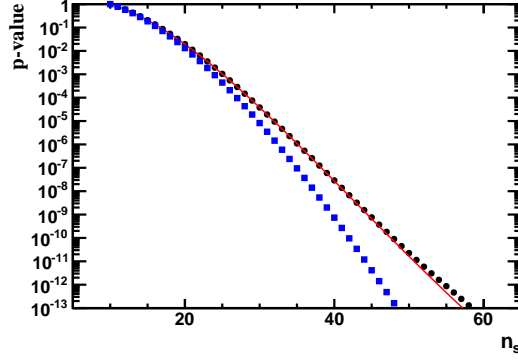


Figure 11. p -values obtained from the asymptotic formula for the profile likelihood as maximized (black points) numerically using MINUIT and (red line) using the analytic approximation. This example used $n_b = 50$, $x = 5$ and $\sigma_y/y = 0.1$. The (blue squares) show the p -values for $\sigma_y = 0$ (no uncertainty in the scaling between regions).

statistical uncertainty in the background rate. In my tests I find that for $\sigma_y/y = 0.1$ this holds for n_b up to about 200.

For the case where $\sigma_y/y > 1/\sqrt{n_b}$, the analytic approximation given is not valid so the values $\hat{s}, \hat{y}, \hat{b}$ must be found numerically, *e.g.*, using MINUIT. This is a valid approach but increases the CPU time required by a factor $\mathcal{O}(100)$. In such cases, however, n_b is large enough that the Poisson term for n_b can be replaced by a Gaussian term. The likelihood is then given by

$$L(n_s, n_b, x|s, b) = \mathcal{P}(n_s, s + b) \cdot \mathcal{G}(b, n_b/x, \sigma_b), \quad (\text{A.10})$$

where the statistical and scale factor uncertainties on the background rate are now included in a single term $\sigma_b^2 = (n_b/x^2)(1 + n_b\sigma_y^2/x^2)$. For large n_b this reduces to $\sigma_b = (n_b/x^2)\sigma_y$. The likelihood in this case can again be maximized analytically giving the following results: $(\hat{s}, \hat{b}) = (n_s - n_b/x, n_b/x)$ and for $s = 0$

$$\hat{b}(0) = \frac{1}{2} \left(n_b/x - \sigma_b^2 + \sqrt{(\sigma_b^2 - n_b/x)^2 + 4\sigma_b^2 n_s} \right). \quad (\text{A.11})$$

Thus, it is possible to provide analytical solutions for all cases.

When searching for a new particle the physical region is $s \geq 0$ and the test statistic used for discovery is $\Lambda(s = 0)$ if $\hat{s} \geq 0$. If $\hat{s} < 0$, I take $\hat{s} = 0$ which gives $\Lambda = 1$. One would expect this to happen at half of all $m(\text{test})$ values considered which is handled naturally by the pseudo-experiment method when obtaining global p -values. When setting limits, I use the *bounded* method from Ref. [7] where the increase in the likelihood is taken from $s = 0$ instead of \hat{s} for the case where $\hat{s} < 0$. This produces limits that are slightly more conservative but also never produces unphysical limits.

⁸This requires differentiating $\log L$ with respect to s and b , setting these to zero, then solving the system of equations.

B. Resonances

This appendix considers the relationship between deviations from the local-linear approximation due to resonance contributions. Figure 12 shows the expected deviations from local-linear for a resonance with $\Gamma/\sigma(m) = 20$. For this case, even if the resonance makes up close to 100% of the total PDF the choice $x = 1$ is still local-linear to about 10%. For smaller resonance contributions larger values of x are local-linear at this level. Figure 13 and 14 show similar plots for $\Gamma/\sigma(m) = 10$ and $\Gamma/\sigma(m) = 5$. For the case $\Gamma/\sigma(m) = 10$, the local-linear approximation is valid at the 10% level up to resonance contributions of about 50% of the total PDF, while for $\Gamma/\sigma(m) = 5$ it is only valid at this level for small resonance contributions.

These results are not surprising. For the case $\Gamma/\sigma(m) = 5$, the signal region (which is $\pm 2\sigma(m)$) contains almost half of the resonance probability. Any large contribution from such a resonance will need to be vetoed. Contributions from wide resonances, however, are *safe* even if they make up the entire PDF. For resonances with widths in the range $20 < \Gamma/\sigma(m) < 5$, applying a veto of the region $|m - m(\text{resonance})| < \Gamma$ will typically be safe for any size resonance contribution. However, if it is known (or can be shown) that the resonance is not dominant, then such a veto may not be required.

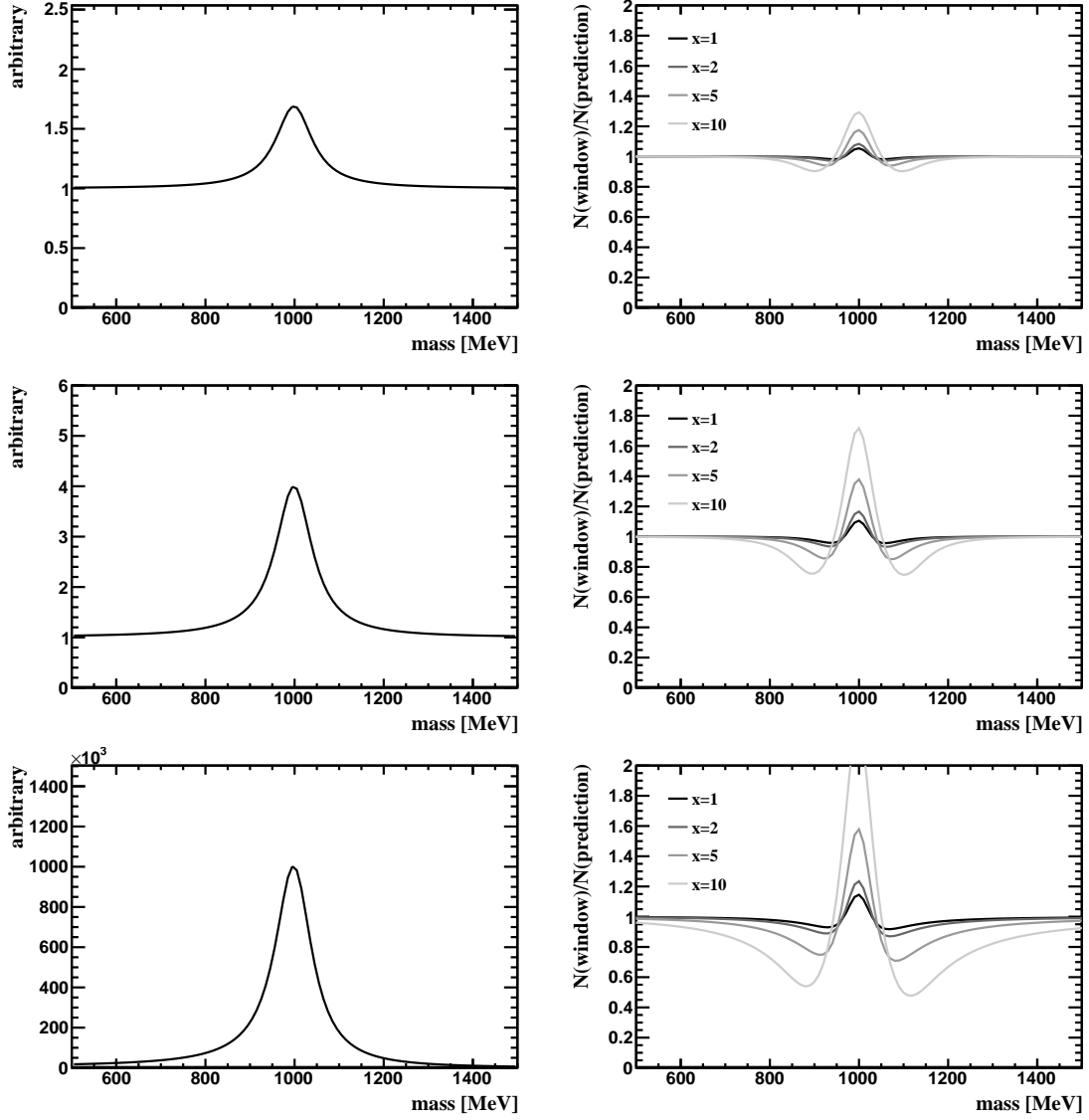


Figure 12. Deviations from local-linear for a resonance with $m = 1000$ MeV, $\Gamma = 100$ MeV, where $\sigma(m) = 5$ MeV. The left plots show the yield distribution (in arbitrary units) for various choices of resonance contribution fraction. The right plots show the ratio of events expected in a signal region to the prediction from the sidebands for various test masses (x -axis) and for various sideband to signal region size ratios x .

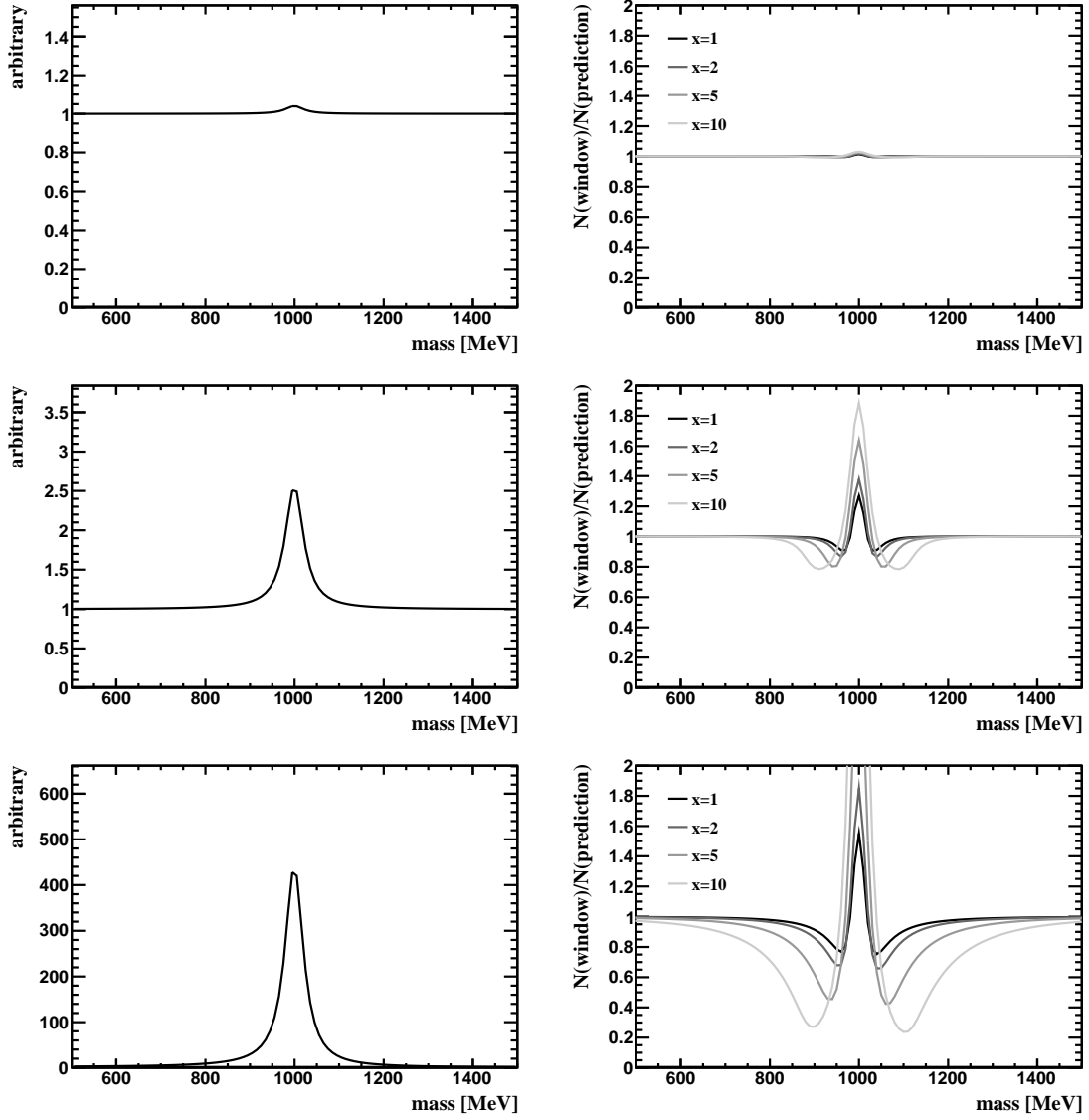


Figure 13. Deviations from local-linear for a resonance with $m = 1000$ MeV, $\Gamma = 50$ MeV, where $\sigma(m) = 5$ MeV. The left plots show the yield distribution (in arbitrary units) for various choices of resonance contribution fraction. The right plots show the ratio of events expected in a signal region to the prediction from the sidebands for various test masses (x -axis) and for various sideband to signal region size ratios x .

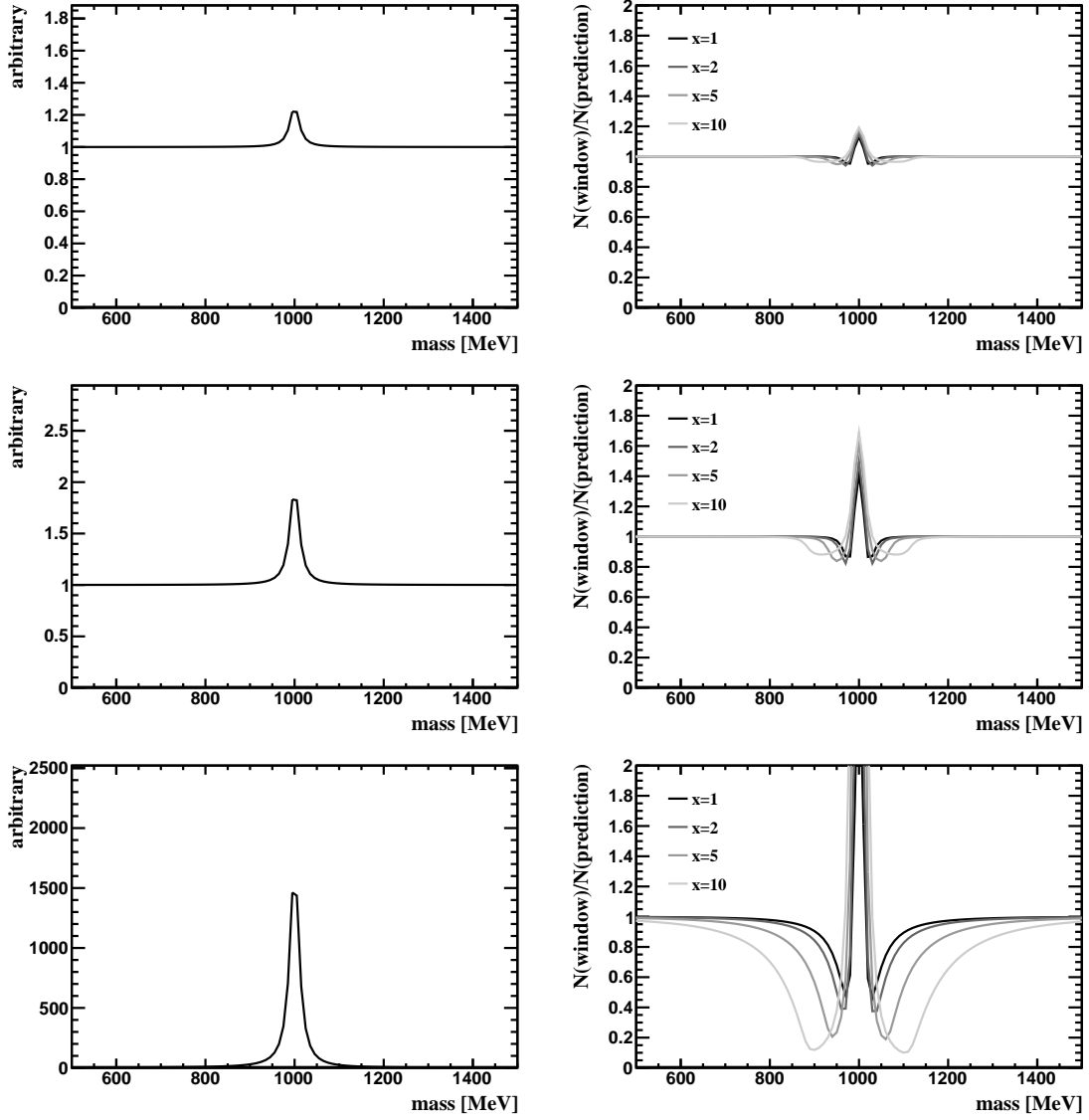


Figure 14. Deviations from local-linear for a resonance with $m = 1000$ MeV, $\Gamma = 25$ MeV, where $\sigma(m) = 5$ MeV. The left plots show the yield distribution (in arbitrary units) for various choices of resonance contribution fraction. The right plots show the ratio of events expected in a signal region to the prediction from the sidebands for various test masses (x-axis) and for various sideband to signal region size ratios x .

C. Two-Sample Non-Parametric Goodness-of-Fit Tests

Assuming that the lifetime distributions of all non-signal PDFs are the same in the signal and background regions, a two-sample non-parametric goodness-of-fit test to the hypothesis that the τ PDF for the signal and background data is the same can be used to test whether a new particle with a resolvable lifetime is contributing to the data. A well-known test is the Kolmogorov-Smirnov (KS)[10] test which uses as test statistic

$$T = \text{MAX}|F_s(\tau) - F_b(\tau)|, \quad (\text{C.1})$$

where F denotes the cumulative distribution. The asymptotic formula is used in this study to obtain an approximate p -value.

Despite the KS test's popularity in particle physics, it is known to not be as powerful as the Cramer-von-Mises (CvM)[11] and Anderson-Darling (AD)[12] tests under many conditions. The CvM and AD tests build statistics

$$T_{\text{CvM}} = \frac{n_s n_b}{n_s + n_b} \int (F_s(\tau) - F_b(\tau))^2 dF_{s+b}(\tau), \quad (\text{C.2})$$

and

$$T_{\text{AD}} = \frac{n_s n_b}{n_s + n_b} \int \frac{(F_s(\tau) - F_b(\tau))^2}{F_{s+b}(\tau)(1 - F_{s+b}(\tau))} dF_{s+b}(\tau), \quad (\text{C.3})$$

respectively. Both are based on the cumulative distributions like the KS test but instead of using only the maximum discrepancy, they use an integrated discrepancy. The AD test provides more weight to the “tails”. The approximate p -values for these tests are obtained using toy simulated data sets in this study.

These two-sample tests are for shape only (in how they are used in this study). To test both size and shape two tests are run: (1) either the KS, CvM or AD tests for a shape comparison of sidebands to search window and (2) Poisson profile likelihood for yield comparisons. As the size and shape tests are uncorrelated (to very good approximation), Fisher's method[13] is used to combine the two p -values to get a single p -value. In this way both shape and size anomalies are tested and a single p -value is obtained.

D. Lifetime Test Results

Figure 15 shows the results of a study of lifetime-only based testing for a given test mass. Pseudo-datasets are generated using a Gaussian τ PDF for prompt backgrounds, and an exponential with effective lifetime of $10\sigma(\tau)$ for displaced backgrounds. Signal events are also generated using an exponential PDF (of varying choices of τ) convolved with a Gaussian to mimic the resolution. The parameters for the background are taken to be $\langle n_s^{\text{prompt}} \rangle = 10$, $\langle n_s^{\text{displ}} \rangle = 0.1$ and $x = 5$. In each case 10 signal events are added. The following tests are studied:

- For comparison an “optimal” test is run where the true background PDF (including for displaced backgrounds) is used. This is a *cheat* since I assume that the displaced background PDF is unknown. The signal PDF is used but with τ as a free parameter.
- Pure shape-based two-sample tests (see Appendix C) are shown in the top left panel. There is not much difference between the KS, CvM and AD tests for this particular data set. As expected these tests provide no power for $\tau \ll \sigma(\tau)$ and increase in power with increasing τ .
- Profile likelihood tests are shown in the top right panel. The Λ test ignores lifetime information (hence its performance is independent of τ). As expected this test is optimal for $\tau \ll \sigma(\tau)$. The Λ_2 test is a two-region (prompt and displaced) counting experiment, where the likelihoods from each region are combined (via a product as usual). This simple test is nearly optimal except when $\tau \sim \sigma(\tau)$. Adding a third region does not improve the performance for this particular displaced background (it may if the displaced background had some additional structure).
- The bottom left panel shows the results of performing the profile likelihood tests along with the shape-based KS test. This greatly enhances the performance of the KS test.
- The bottom right panel, however, shows that performing the KS test with the two-region profile likelihood test only provides a small gain in the region $\tau \sim \sigma(\tau)$.

While the combination of KS and Λ_2 in theory improves the performance for $\tau \sim \sigma(\tau)$, it adds some complexity to the analysis. For example, an additional assumption has now been employed: that all lifetime PDFs are the same in the signal and sideband regions. This is likely to be true for the prompt SM background; however, it may not be true for other types of background. Furthermore, the gain in significance obtained by increasing the number of generated signal events from 10 to 11 for $\tau \sim \sigma(\tau)$ is much greater than that obtained by performing the KS test along with the profile likelihood; *i.e.*, the maximal gain in sensitivity in this example of using both the KS and Λ_2 tests is $< 10\%$ in signal rate sensitivity. This gain requires adding an assumption which may not be valid and is expected to be difficult to validate/study. Thus, my conclusion is that the nominal test should simply be Λ_2 .

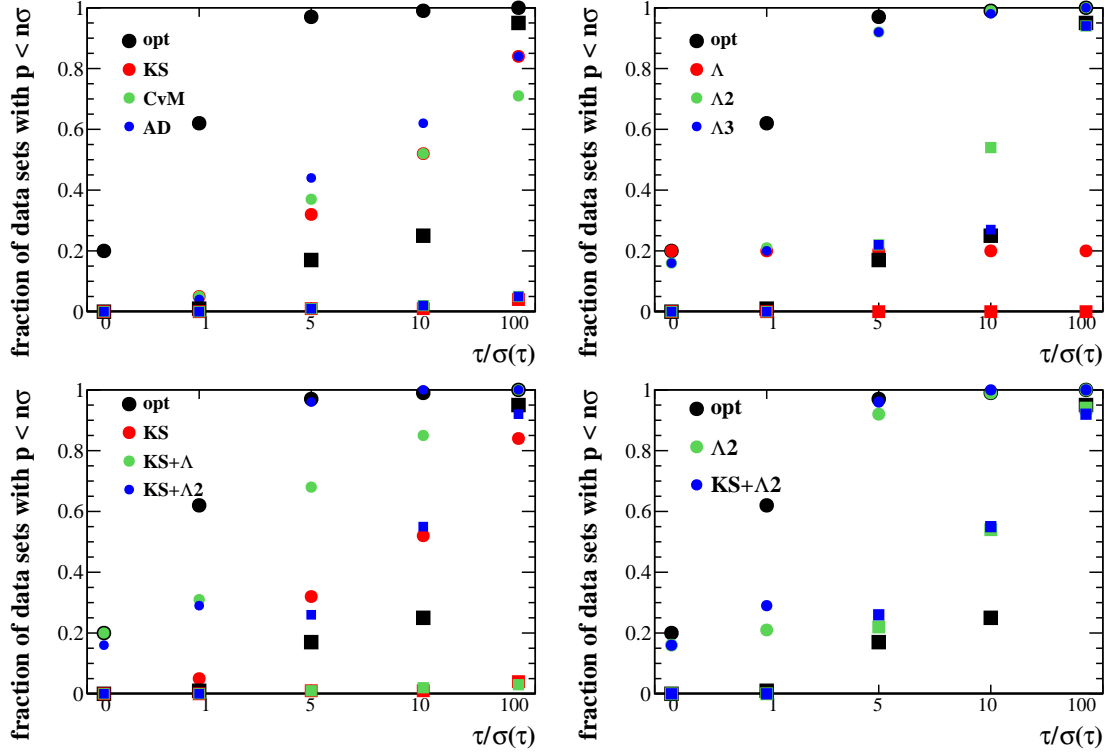


Figure 15. The (circles,squares) represent the fraction of data sets with $p < (3,5)\sigma$. The test labels are defined as follows: (opt) the *cheat* PDF-based likelihood fit; (KS,CvM,AD) the 2-sample tests (see text); ($\Lambda(2,3)$) the profile likelihood test, including the (2,3)-region versions; ($KS + \Lambda(2)$) combination of both the KS and profile likelihood tests. *N.b.*, the size of the markers has no meaning; this was done to aid in viewing markers for tests with similar power.

E. Fast Algorithm

The two-region profile likelihood method can be made extremely fast. In Appendix A I obtained analytic expressions that maximize the likelihood (including the case where a Gaussian uncertainty is included on the scaling factor that relates the sideband yields to those in the signal region); thus, no numerical minimization routine, *e.g.*, MINUIT, needs to be run. Note that the data can be binned in mass (in both the prompt and displaced regions) to perform this test. This means that time consuming event loops are not required. Furthermore, as one scans the mass range, rather than summing up the yields in the signal and background regions, it is faster to simply subtract the bin(s) removed from each region and then add only the bin(s) added. Combining all of these optimizations results in a test that can be performed on 10M data sets with $\mathcal{O}(1000)$ test masses in about 2 hours on a single CPU core. This process is trivially parallelized to run on multiple cores. Therefore, it is possible to confirm a significance of $> 5\sigma$ without the need to rely on an asymptotic p -value. This is a desirable feature since for rare decays the asymptotic formulae tend to underestimate the significance for very low p -values.